Assignment 1 - CSC/DSC 265/465 - Spring 2019 - Due February 12, 2019

Unless otherwise specified, statistical significance can be taken to hold when the relevant P-value is no larger than $\alpha = 0.05$. Note that problem **Q4** is reserved for graduate students.

Q1: For this question, use the `OJ` data set from the `ISLR` package (you can install this package from the `CRAN` repository directly from `R`). This represents data from sales of two brands of orange juice. Each of the $n = 1070$ observations represents a single sales transaction. We will make use of the variables:

- `Purchase` = Purchased brand was either *Citrus Hill* (`= CH`) or *Minute Maid* (`= MM`).
- `StoreID` = ID of store at which purchase was made (`StoreID = 1,2,3,4,7`).
- `LoyalCH` = Customer brand loyalty score for `CH` on a scale of 0 to 1.

The objective is to determine whether or not customer loyalty differs significantly between stores.

(a) Construct side-by-side boxplots of `LoyalCH` using `Purchase` as the group variable. Interpret what you see. Use a Wilcoxon rank sum test to determine if there is a significant difference in the median of `LoyalCH` score between purchase groups.

(b) Construct side-by-side boxplots of `LoyalCH` using `StoreID` as the group variable. Fit an ANOVA model using `LoyalCH` as response and `StoreID` as the treatment variable. Is there evidence that mean loyalty score varies by store? (At this point, you need not consider any transformation of the response variable).

(c) Using Tukey's pairwise procedure, what can be said about the rankings of the mean loyalty scores, using a family-wise error rate of $\alpha_{FWE} = 0.05$. You can use function `TukeyHSD`.

(d) Noting that `LoyalCH` is constrained to be between 0 and 1, it is important to assess whether or not the distributional properties of the responses permit the reporting of accurate observed significance levels. Construct a normal quantile plot of the residuals from your ANOVA fit. Then apply the empirical rule to assess the normality of the residuals. What do you conclude? Why should this be done using the model residuals instead the response variable directly?

(e) We can use simulation methods to judge the accuracy of the observed significant levels. Suppose the true mean value μ of `LoyalCH` does not vary by store. Then a response from any store can be modeled as $y = \mu + \epsilon$, where ϵ is a zero mean error term. The distribution of ϵ can then be estimated using the residuals.

We can do this using a *bootstrap* procedure (Section 10.2 of lecture notes, Section 5.2 of ISLR). Suppose `y` is the response vector of length n, and `x` is the factor variable identifying the store. We have already fit the model `y` $\sim$ `x`. Suppose we then let `y.boot` be a random sample of size n (with replacement) of the residuals from some fitted model. This is equivalent to simulating a sample from $y = \mu + \epsilon$, where $\mu = 0$. This suffices for our purpose, since the actual value of μ will play no role in the procedure (and so can be zero with no loss of generality).

If we then fit the ANOVA model `y.boot` $\sim$ `x`, the null hypothesis of equal treatment means will hold, therefore the P-value of the F-test should possess a uniform distribution on $[0, 1]$. Of course, this depends on the correctness of the distributional assumptions (that ϵ is normally distributed), and so provides a means of assessing whether or not those hold (or at least that any deviation from normality does not significantly affect the accuracy of the reported level of significance).

To carry out the procedure, simulate M bootstrap samples, capturing the P-value from the F-test for each one. If the distributional assumptions required for the F-test hold, then the replicated P-value distribution should be approximately uniform.

(i) Suppose X is a continuous random variable with CDF $F(x) = P(X \leq x)$. Verify that $F(X)$ and $1 - F(X)$ have a uniform distribution on $[0, 1]$. How does this verify the claim made above that the P-value is uniformly distributed under the null hypothesis?

(ii) Carry out this bootstrap procedure for the ANOVA model fit in Part (b). Use $M = 100,000$. Draw a histogram of the replicated P-values, using the `nclass = 25` option. Report the proportion of the replicated P-values, say $\hat{\alpha}$, below $\alpha = 0.001, 0.01, 0.05, 0.1$. In addition, for each value of α, report $Z = (\hat{\alpha} - \alpha)/SE$, where $SE = \sqrt{\alpha(1-\alpha)/M}$ is the standard error of $\hat{\alpha}$. Interpret your results.

(iii) What is the standard error of $\hat{\alpha}$ for $\alpha = 0.1$ and M? What does this tell you about the overall accuracy of the bootstrap procedure.

(f) We will next carry out an experiment to assess the sensitivity of the bootstrap method. Consider the transformation $y^* = 1/(1-y)$, and apply it to the responses of store `StoreID == 7`. Repeat the bootstrap procedure just described, except that the replicated response vector `y.boot` will be constructed by sampling n responses with replacement from the transformed responses y^*. Note that although we only sample responses from store `StoreID == 7`, responses for all original stores are being simulated. Would the observed significance levels using data with this distribution be accurate?

(g) Repeat Part (f), but apply a Box-Cox transformation to the replicated samples (see Section 10.8.3 of the CSC/DSC 462 lecture notes in the `COURSE MATERIALS` folder on the Blackboard course website). You will need to load the `MASS` package. Under this transformation, would the observed significance levels be accurate?

Q2: We will add a new variable to the analysis of problem **Q1**, namely `PriceDiff` (sale price of MM less sale price of CH).

(a) Fit the linear model

$$\texttt{LoyalCH} = \beta_0 + \beta_1 \times \texttt{PriceDiff}. \text{ [Model 1]}$$

Is there significant evidence that `LoyalCH` varies with `PriceDiff`? If so, in what way?

(b) Fit the linear model (or one equivalent to):

$$\texttt{LoyalCH} = \beta_0 + \beta_1 \times I\{\texttt{StoreID == 1}\} + \ldots + \beta_4 \times I\{\texttt{StoreId == 4}\}. \text{ [Model 2]}$$

How does this compare to the ANOVA model of **Q1** Part (b)? **HINT:** Using the `lm` function, it is not necessary to construct individual indicator variables. Using `R` formula objects, this is done automatically if a formula such as `y` $\sim$ `x.factor` is used, where `x.factor` is a vector of `factor` type.

(c) We next combine the two predictor variables. This can be done additively (using `R` formula notation):

$$\texttt{LoyalCH} \sim \texttt{PriceDiff} + \texttt{StoreID}, \text{ [Model 3]}$$

or by including all interactions:

$$\texttt{LoyalCH} \sim \texttt{PriceDiff} * \texttt{StoreID}. \text{ [Model 4]}$$

For each of the four models, superimpose these fits graphically on a scatterplot of the original data. For Models 2,3,4 present clearly the regression line separately for each store. **HINT:** One way to do this is to use the `newdata` option in the `predict` method. Set `newdata` equal to a data frame with columns `PriceDiff` and `StoreID`. For each of the five levels of `StoreID` include two rows, with `PriceDiff` equal to the endpoints of the range of `PriceDiff`. After some rearranging of the output of `predict`, the `matplot` function can be used to draw 5 separate lines on a single plot. In addition, this same method can be used for all models.

(d) Construct a table giving the SSE and the SSE (residual) degrees of freedom for each model. Which models are nested? Do a goodness of fit test to determine if Model 3 improves Model 1 or 2, and if Model 4 improves model 3. Do this directly using the SSE values. Then verify your result using the `anova` function.

(e) We next consider the possibility that `LoyalCH` varies with `PriceDiff` within some but not other stores. We can do this separately for, say, `StoreID == 1` by adding the term

$$\beta' \times \texttt{I(PriceDiff*(StoreID==1))}$$

to Model 2. Do this for each `StoreID` level, and report a P-value for the F-test comparing the full and reduced model.

(f) If you apply the Bonferroni multiple test procedure to the output of Part (e), with a family-wise error rate of $\alpha_{FWE} = 0.05$, for which stores is there evidence that `LoyalCH` varies with `PriceDiff`?

Q3: We are given a multiple linear regression model

$$y_i = \beta_1 x_{i1} + \ldots + \beta_q x_{iq} + \epsilon_i, \;\; i = 1, \ldots, n,$$

where ϵ_i are *iid* error terms with $\epsilon_i \sim N(0, \sigma^2)$. Let $\hat{\beta}_i$ be the least squares estimate of β_i, $i = 1, \ldots, q$. There may be an advantage with respect to the interpretability of the regression coefficients if they are uncorrelated. In this case, the contribution of each predictor to the model can be assessed independently of the other predictors.

(a) If $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ldots \hat{\beta}_q]^T$ is the vector of least squares coefficient estimates, then the covariance matrix is given by

$$\Sigma_{\hat{\boldsymbol{\beta}}} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}. \tag{1}$$

Using this expression, show that the regression coefficients are mutually uncorrelated if and only if

$$\sum_{i=1}^{n} x_{ij} x_{ik} = 0 \tag{2}$$

for each pair $j \neq k$. If the columns of $\mathbf{X}$ are *orthogonal*, $\mathbf{X}^T\mathbf{X}$ will be a diagonal matrix, and if the columns are *orthonormal*, $\mathbf{X}^T\mathbf{X}$ will be the identity matrix.

(b) Polynomial regression extends simple linear regression by constructing new predictor variables from powers of a single predictor variable $\mathbf{x}$:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p. \tag{3}$$

Then we have

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^p \end{bmatrix}$$

The problem with this approach is that $\mathbf{X}$ may possess significant collinearity (Section 6.6 of lecture notes, Section 3.3.3 of ISLR). One solution is to use as predictors the linear transformation

$$\mathbf{X}' = \mathbf{X}\mathbf{A} = \begin{bmatrix} z_{10} & z_{11} & z_{12} & \ldots & z_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n0} & z_{n1} & z_{n2} & \ldots & z_{np} \end{bmatrix}$$

where $\mathbf{A}$ is a $(p+1) \times (p+1)$ matrix of coefficients. Then $\mathbf{A}$ can be chosen so that $\mathbf{X}'$ is orthogonal. Note that the first column of $\mathbf{X}'$ has been relabelled $j = 0$. The coefficient matrix $\mathbf{A}$ is usually upper triangular, so the transformed predictor variables become

$$\begin{aligned} z_{i0} &= a_{00} \\ z_{i1} &= a_{01} + a_{11} x_i \\ z_{i2} &= a_{02} + a_{12} x_i + a_{22} x_i^2 \\ &\vdots \\ z_{ip} &= a_{0p} + a_{1p} x_i + a_{2p} x_i^2 + \ldots + a_{pp} x_i^p \end{aligned}$$

noting that the first row and column of $\mathbf{A}$ are here labeled $i = j = 0$. The model is now

$$y_i = \beta'_0 z_{i0} + \beta'_1 z_{i1} + \beta'_2 z_{i2} + \ldots + \beta'_p z_{ip}. \tag{4}$$

Show that the two models are equivalent in the sense that the fitted values $\hat{y}_i$ must be the same. How are the least squares estimates of the coefficients for the respective models related?

(c) Suppose we wish to choose $\mathbf{A}$ so that the components of $\hat{\boldsymbol{\beta}} = [\hat{\beta}_0 \ldots \hat{\beta}_p]^T$ are uncorrelated. Verify that this is achieved by the choice

$$\mathbf{A} = \begin{bmatrix} 1 & -\bar{x} \\ 0 & 1 \end{bmatrix}$$

for $p = 1$, where $\bar{x} = n^{-1} \sum_i x_i$.

(d) The function `poly` can be used to construct matrices of the form $\mathbf{X}$ or $\mathbf{X}'$ (note that the first column of one's is not included). Let `x` be a vector of length n representing a single predictor variable. Then `poly(x, 3, raw=TRUE)` produces a matrix of the form $\mathbf{X}$ (that is, the jth column is simply the jth powers of `x`). On the other hand, `poly(x, 3, raw=FALSE)` produces a matrix of the form $\mathbf{X}'$, where the coefficients are chosen so that the columns are orthonormal (ie. $[\mathbf{X}']^T\mathbf{X}' = I$, where I is the $p \times p$ identity matrix).

Generate simulated data for a linear model with the following code. Here `x` is the predictor variable and `y` is the response.

```
> set.seed(12345)
> x = (1:100)/100
> y = rnorm(100,mean=1+5*x^2,sd=1)
> plot(x,y)
```

(i) Write the model explicitly, describing the error term precisely.

(ii) Fit the model twice, using the methods given next:

```
> fit1 = lm(y~poly(x,3,raw=T))
> fit2 = lm(y~poly(x,3,raw=F))
```

For each fit report and interpret the `F-statistic` given by the `summary` method for each fit.

(iii) Construct a scatter plot of the response variable `y` against predictor variable `x`. Superimpose the fitted values of both fits. Make sure you use distinct plotting characters and colors, so that the two sets of fitted values can be distinguished. For example, you could use the `matplot` function with options `pch=c(2,3)`, `col=c('green','red')` and `add=T`.

(iv) Each fit summary reports P-values for rejecting the null hypothesis $H_0 : \beta_j = 0$ for each coefficient. Compare the P-values for the two fits. Do either summaries permit the (correct) conclusion that the mean of the response y_i is a second order polynomial in the predictor variable x_i? If for either fit the P-values for the coefficients (other than the intercept) are all large (ie. > 0.05) does this contradict the conclusion of Part (e)-(ii)?

Q4: [For Graduate Students] Consider the matrix representation of the multiple linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is an $n \times 1$ response vector, $\mathbf{X}$ is a $n \times q$ matrix, $\boldsymbol{\beta}$ is a $q \times 1$ vector of coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of error terms. The least squares solution is expressed using the coefficient vector $\boldsymbol{\beta}$ which minimizes the error sum of squares

$$SSE[\boldsymbol{\beta}] = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

(a) By setting each partial derivative $\partial SSE[\boldsymbol{\beta}]/\partial\beta_j$ to zero, $j = 1, \ldots, q$, verify that the least squares solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

(b) Next, suppose we wish to minimize $SSE[\boldsymbol{\beta}]$ subject to $m < q$ linear constraints on $\boldsymbol{\beta}$, expressed in matrix form as

$$\mathbf{C}\boldsymbol{\beta} - \mathbf{d} = \mathbf{0} \tag{5}$$

where $\mathbf{C}$ is an $m \times q$ matrix and $\mathbf{d}$ is an $m \times 1$ column vector and $\mathbf{0}$ is an $m \times 1$ column vector of zeros. This is equivalent to the m linear constraints

$$\begin{aligned} c_{11}\beta_1 + \ldots + c_{1q}\beta_q &= d_1 \\ &\vdots \\ c_{m1}\beta_1 + \ldots + c_{mq}\beta_q &= d_m, \end{aligned}$$

where c_{ij} and d_k are the respective elements of $\mathbf{C}$ and $\mathbf{d}$. Using the Lagrange multiplier method, we can calculate the *constrained least squares* solution $\hat{\boldsymbol{\beta}}_c$ to this problem by minimizing

$$\Lambda = SSE[\boldsymbol{\beta}] + \bar{\lambda}^T(\mathbf{d} - \mathbf{C}\boldsymbol{\beta})$$

with respect to $(\boldsymbol{\beta}, \bar{\lambda})$, where $\bar{\lambda}^T = (\lambda_1, \ldots, \lambda_m)$, while applying constraint (5). Verify that

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}}_u + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\left[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\right]^{-1}(\mathbf{d} - \mathbf{C}\hat{\boldsymbol{\beta}}_u),$$

where $\hat{\boldsymbol{\beta}}_u = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is the unconstrained least squares solution.